Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Exp Psychol Gen. Author manuscript; available in PMC 2024 June 01.

Published in final edited form as:

J Exp Psychol Gen. 2023 June ; 152(6): 1565–1579. doi:10.1037/xge0001336.

Collaborative decision making is grounded in representations of other people's competence and effort

Yang Xiang[1,*], **Natalia Vélez**[1], **Samuel J. Gershman**[1,2,3]

[1]Department of Psychology, Harvard University

[2]Center for Brain Science, Harvard University

[3]Center for Brains, Minds and Machines, MIT

Abstract

By collaborating with others, humans can pool their limited knowledge, skills, and resources to achieve goals that outstrip the abilities of any one person. What cognitive capacities make human collaboration possible? Here, we propose that collaboration is grounded in an intuitive understanding of how others think and of what they can do—in other words, of their mental states and competence. We present a belief-desire-competence framework that formalizes this proposal by extending existing models of commonsense psychological reasoning. Our framework predicts that agents recursively reason how much effort they and their partner will allocate to a task, based on the rewards at stake and on their own and their collaborator's competence. Across three experiments (N = 249), we show that the belief-desire-competence framework captures human judgments in a variety of contexts that are critical to collaboration, including predicting whether a joint activity will succeed (Experiment 1), selecting incentives for collaborators (Experiment 2), and choosing which individuals to recruit for a collaborative task (Experiment 3). Our work provides a theoretical framework for understanding how commonsense psychological reasoning contributes to collaborative achievements.

1 Introduction

Collaboration enables humans to achieve goals beyond the capabilities of any one individual: No one can build a city, land on the moon, or even carry a couch up a flight of stairs entirely on their own. Humans are unique in their ability to form collaborative arrangements that are maintained over generations, as in the case of institutions, and that span continents, as in the case of large scientific collaborations. Compared to other animals, we've made it a key part of our niche (Tomasello et al., 2005), and the roots of this ability arise early in development (see Tomasello and Hamann 2012, for a review). How do we do it? What cognitive capacities give rise to these collaborative achievements?

[*]Corresponding author: yyx@g.harvard.edu.

Y.X., N.V., and S.J.G. conceived the project. Y.X. implemented the experiments and analyzed the data. Y.X., N.V., and S.J.G. wrote the manuscript and approved the final version for submission. A preprint of the manuscript has been posted on PsyArXiv: https://psyarxiv.com/gcnrq/.

Data and code used for simulation and analysis are available at https://github.com/yyyxiang/competence_effort.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

A key part of the answer is that, in order to work with others, we have to understand how our collaborators think and act. Philosophical treatments of collaboration provide an inventory of basic mental states and commitments required for collaborative action. To fulfill these criteria, agents must hold certain representations in mind, including a representation of shared goals, knowledge and beliefs about which actions will lead to the fulfillment of the goal, and intentions to carry out said actions (Pollack, 1990; Searle, 1990; Grosz and Kraus, 1996; Bratman, 1992). Recent work suggests that agents plan their actions by anticipating those of their collaborative partners, revealing that people minimize joint action costs (Török et al., 2019, 2021), compute a "mind of the group" (i.e., an average group member's mind; Khalvati et al., 2019), as well as represent, monitor, and predict each other's actions to achieve a joint goal (Sebanz and Knoblich, 2021; Vesper et al., 2010; Wu et al., 2021). Thus, fundamentally, successful collaboration requires that we understand how others think and act.

Bayesian models of commonsense psychological reasoning formalize these computations as theory of mind (Premack and Woodruff, 1978), which takes the form of belief-desire reasoning: People act based on their beliefs to achieve their desires (Baker et al., 2009, 2017). However, even state-of-the-art psychological models are missing two ingredients that are particularly important to successfully work together. The first missing ingredient is a representation of one's own and of others' competence. A person's competence entails all kinds of physical and mental abilities, such as strength, speed, memory, linguistic skill, etc. Traditional models of belief-desire reasoning assume that agents are capable of fulfilling their desires as long as the constraints of the environment allow it. But this is not always so—an individual might have the necessary beliefs and desires to complete a task, but lack the competence to carry it out. For example, imagine a child who knows where a box is, has the desire to lift the box for a cookie as reward, but is not strong enough do so. The child would not be able to lift the box and get the cookie reward herself. However, she could choose to collaborate with another child who is strong enough to lift the box but does not know where it is. This example illustrates the necessity of reasoning about one's own and other people's competence in collaborative activities. Past work suggests that even young children are capable of inferring other people's competence based on the difficulty of the tasks they are willing to take on (Jara-Ettinger et al., 2015, 2016). However, models of commonsense psychological reasoning have yet to incorporate a representation of other people's competence that is separate from other costs, such as the inherent difficulty of a task.

The second missing ingredient is a formal theory of how individuals allocate effort to joint tasks. Recent work suggests that even young children use social information to decide how to allocate effort to a task; for example, young children might observe others to infer the difficulty of a task (Lucca et al., 2020) or use their example to determine whether expending effort will pay off (Leonard et al., 2017, 2020). Children are also capable of dividing physical and cognitive labor in collaborative tasks based on relative ability and task difficulty (Magid et al., 2018; Baer and Odic, 2022). More recent models have enriched theories of psychological reasoning by demonstrating that people do not merely expect others to fulfill desires—thus solely maximizing rewards—but rather expect others to act as utility maximizers—balancing the rewards of an outcome against the costs of achieving it

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(Jara-Ettinger et al., 2016). However, existing models do not capture the fact that engaging in collaborative action reduces the effort that any one individual needs to expend on a task: lifting a couch alone may take an intense all-out effort, but feel significantly easier when the burden is shared between two friends. It remains an open question how agents figure out how to adjust their effort in collaborative settings—for example, how we anticipate the effort of others and calibrate our own in order to lift the couch successfully without tiring ourselves completely.

To fill these gaps, we present a model of belief-desire-competence reasoning that extends existing models of commonsense psychological reasoning. Our model predicts that agents recursively reason how much effort they and their partner will allocate to a task, based on the rewards at stake and on their own and their collaborator's competence. In addition, we contrast our model with three alternative models that use simpler heuristics (solitary, compensatory, and maximum effort models) and show that joint reasoning is necessary for explaining collaborative behavior. The alternative models might seem impossible at first blush, but they are not a priori impossible for all the scenarios. Moreover, the alternative models are motivated by phenomena such as social compensation and social loafing (see Theoretical Framework, below). Therefore, demonstrating that these models are not sufficient to explain behavior in our studies provides support for our claim that collaborative decision making is grounded in joint reasoning about competence and effort.

Across three experiments ($N = 249$), we demonstrate that the belief-desire-competence framework captures human judgments in a variety of contexts that are critical for collaboration. In each experiment, participants watched contestants in a game show attempt to lift a box to win cash prizes. In Experiment 1, each contest was divided into three rounds: In the first two rounds, contestants attempted to lift the box on their own, in order to provide information about their strength and about how they respond to incentives. In the third, critical round, two contestants then attempted to lift the box together. Participants inferred how strong each contestant is, how much effort they would put into lifting the box in the third round, and how likely they are to succeed. In Experiments 2–3, participants saw contestants with different strengths who required different incentives to lift a heavy weight, and they decided how much incentive to provide (Experiment 2) and which contestants to recruit (Experiment 3) to lift boxes of different weights. Across all models, we find that a model that recursively infers how much effort to allocate best matches human judgments, compared to various alternative models that implement simpler heuristics rather than reasoning recursively about effort. Together, our results provide evidence that collaborative decisions are grounded in an intuitive understanding of how others think and of what they can do.

2 Theoretical Framework

We propose a belief-desire-competence framework (Figure 1) to illustrate how people's reasoning about beliefs, desires, and competence determine the action they take. After laying out the framework, we formalize the idea in a simple context of box-lifting.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2.1 Overview of the belief-desire-competence framework

We assume that the agent's desires and beliefs about the internal state and external environment determine the amount of effort the agent would exert. The action to take depends on both the agent's effort and competence, which encompasses all kinds of physical and mental abilities. The final outcome is a function of the difficulty of the task and the action the agent takes. The agent updates their beliefs after observing the outcome.

We simplified the framework and formalized it in a setting where agents of different strengths attempt to lift boxes of different weights. This allowed us to generate hypotheses to test the key predictions of this framework. Specifically, the agent's desire is the reward they get from lifting the box, the agent's competence is their physical strength, and the task difficulty is the box weight. Observing the outcome (either Lift or Fail) allows the agent to update their belief about their strength, the box weight, etc. One advantage of testing the model in a physical domain like box lifting, compared to cognitive domains such as numerosity judgments, is that the task difficulty and agents' competence and effort have concrete units. Thus, we can express the relationship between effort and action outcomes in a simple, deterministic model: Agents' competence and effort translate into the force that they exert, and they succeed if that force exceeds the weight of the box. [1] Below we provide a formal description of our model predicated on this simplified framework.

2.2 Formal description

We first consider a single-agent case, then generalize it to a multi-agent setting which involves recursive reasoning between agents, followed by three alternative models without recursive reasoning.

2.2.1 Single-agent model—The single-agent model describes how an agent decides the amount of effort to exert in an attempt to lift a box. It consists of the following components:

- $S \geq 0$ is the strength of the agent, expressed in units of weight (i.e., the maximum amount of weight they are able to lift). We assume strength is a static property of each agent.
- $W \geq 0$ is the weight of the box.
- $E \in [0, 1]$ is the effort the agent puts into lifting the box. It defines the proportion of an agent's strength that is applied to lifting.[2]
- $L = 1$ indicates that the agent lifts the box ($L = 0$ otherwise). This is a deterministic function of effort, strength and weight: $L = \mathbb{I}[E \cdot S \geq W]$, where the indicator function $\mathbb{I}[\cdot] = 1$ if its argument is true, 0 otherwise.
- $R \geq 0$ is the reward for successfully lifting the box.

[1] In real-world lifting events, it is possible for a lifter to fail to lift a weight, even if they are strong enough to lift it and exert the needed effort, due to slight variations in body posture, hand positioning, etc. However, we think that a deterministic function provides a reasonable approximation within this simplified task.

[2] For computational tractability, we constrained the effort space to discrete values, ranging from 0 to 1 with increments of 0.05.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

- $C(E)$ is the agent's effort cost. For simplicity, we adopt a linear cost function $C(E) = \alpha E$, where $\alpha \geq 0$ is a scaling parameter that denotes the agent's laziness (larger α means lazier)[3]
- $U(E)$ is the utility function that measures the net reward minus cost:

$$U(E) = R \cdot P(L = 1 \mid E, W) - C(E), \tag{1}$$

where the probability of lifting the box is given by marginalizing over all possible strengths:

$$P(L = 1 \mid E, W) = \int_S P(S)\mathbb{I}[E \cdot S \geq W]dS\,. \tag{2}$$

This model can be straightforwardly extended to account for uncertainty about W and α.

We model agents as utility maximizers, such that the optimal effort is:

$$E^* = \underset{E}{\operatorname{argmax}} U(E)\,. \tag{3}$$

2.2.2 Multi-agent joint effort model—The multi-agent joint effort model generalizes the single-agent model to a multi-agent setting. The model is termed "joint effort" because agents reason about others' effort recursively when deciding how much effort to allocate. When multiple agents attempt to lift a box together, the lift outcome depends on every agent's effort, strength, and the box weight:

$$L = \mathbb{I}[\sum_a E_a S_a \geq W], \tag{4}$$

where a indexes agents. The optimal effort is given by:

$$\mathbf{E}^* = \underset{\mathbf{E}}{\operatorname{argmax}} R \cdot P(L = 1 \mid \mathbf{E}, W) - C(\mathbf{E}), \tag{5}$$

where $\mathbf{E} = [e_1, \ldots, e_n]$ is the vector of efforts for n agents. In practice, we solve the joint optimization by fixed point iteration: holding the efforts of all but one agent fixed, we maximize the utility with respect to the focal agent, iterating over agents until convergence (convergence threshold set to 0.06).

People have a general preference for fairness (Fehr and Schmidt, 1999; Tabibnia et al., 2008), and these fairness preferences play a role in human collaboration (Blake et al., 2015; Hamann et al., 2011). In light of this, we add to the cost function a Gini coefficient that penalizes unequal effort allocations. [4] The Gini coefficient is a very widely-used measure

[3]While we refer to the α parameter as *laziness*, it can also be understood as a form of competence-dependent reward scaling. For example, an agent might deserve to be paid more due to their superiority over others in terms of competence.

[4]We contrast the joint effort model with and without the Gini coefficient in the supplemental material (Figures S1-S5). From Figures S1B, we can clearly see that model predictions of effort change drastically with and without the Gini coefficient, thereby demonstrating the importance of penalizing unequal effort allocations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of income inequality in economics (Atkinson et al., 1970; Sen et al., 1997; Campano and Salvatore, 2006), and has been applied to a myriad of fields to measure education inequality (Thomas et al., 2001), health inequality (Regidor, 2004), and yield inequality (Sadras and Bongiovanni, 2004). The Gini coefficient is defined as half of the relative mean absolute difference in effort (Sen et al., 1997):

$$G(\mathbf{E}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mid E_i - E_j \mid}{2n\sum_{j=1}^{n} E_j}, \tag{6}$$

where $|E_i - E_j|$ denotes the absolute difference in effort between each pair of agents (i, j). To take fairness preferences into account, we add a term to the cost function:

$$C(\mathbf{E}) = \sum_{a} \alpha_a E_a + \beta G(\mathbf{E}), \tag{7}$$

where $\beta \geq 0$ is a scaling parameter. Note that we allow different agents to have different effort cost coefficients (α_a), an assumption that we explore further in Experiment 3.

2.2.3 Alternative models—In real life, not all teamwork operates as the joint effort model suggests. It is not uncommon that in some group projects, one person does almost all the work while the others do very minimal work, if at all. Indeed, similar types of behaviors have been documented in the literature and explained as *social compensation* (Williams and Karau, 1991), where individuals work hard collectively when they expect their teammates to be less invested or less capable, and *social loafing* (Karau and Williams, 1993; Latané et al., 1979), where free riders take advantage of their teammates and expect them to pick up the slack. Inspired by these phenomena, we introduce two alternative models which do not invoke recursive reasoning of effort: The solitary effort model and the compensatory effort model.

The **solitary effort** model instantiates an extreme case of social compensation. Agents assume that their partners will expend 0 effort on the task, and allocate effort as though they were performing the task alone:

$$\mathbf{E}_{/a} = 0, \tag{8}$$

where */a* indexes all the agents except agent *a.* Each agent's optimal effort only depends on their own strength, effort cost, and the box weight:

$$E_a^* = \underset{E_a}{\operatorname{argmax}} R \cdot \mathbb{I}[E_a S_a \geq W] - \alpha_a E_a . \tag{9}$$

Therefore, we have:

$$\begin{aligned} E_a^* &= \underset{E_a}{\operatorname{argmax}} R \cdot P(L = 1 \mid E_a, \mathbf{E}_{/a} = 0, W) - \alpha_a E_a \\ &= \underset{E_a}{\operatorname{argmax}} U(E_a), \end{aligned} \tag{10}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $U(E_a)$ is the single agent utility function described above (Eq. 1). Note that the solitary effort model does not include a Gini coefficient. This is because a model that instantiates an extreme case of social compensation likely does not care about (un)fairness. The same reasoning applies to the compensatory effort model. [5]

The **compensatory effort** model instantiates the other extreme (social loafing): Agents assume that their partners will exert maximal effort, and therefore the agents will exert only the minimum effort needed to accomplish the task:

$$\mathbf{E}_{/a} = 1\,. \tag{11}$$

The optimal effort is given by:

$$E_a^* = \underset{E_a}{\operatorname{argmax}} R \cdot P(L = 1 \mid E_a, \mathbf{E}_{/a} = 1, W) - \alpha_a E_a\,. \tag{12}$$

The solitary effort and compensatory effort models make the same predictions as the joint effort model for single-agent events, but diverge for multi-agent events. The joint effort model assumes that agents reason recursively about each other's effort, while the solitary effort and compensatory effort models make fixed assumptions about the efforts of other agents.

In addition to the two alternative models described above, we include a third alternative: the **maximum effort** model. This model is motivated by an implicit assumption in some previous research (e.g., Jara-Ettinger and Gweon, 2017) that people expect agents to exhibit their full capacity when they take actions. The maximum effort model simply assumes that agents are always exerting all their effort regardless of their strength, effort cost, partners, etc., which means $\mathbf{E} = 1$. This is similar to the compensatory effort model, except that here agents do not modulate their own effort allocation in response to the maximal effort of their partners.

2.3 Model implementation and assessment

We implemented the model in WebPPL, a probabilistic programming language embedded in Javascript (Goodman and Stuhlmüller, 2014).

The strength samples were drawn from a uniform distribution with a lower bound of 1 and upper bound of 10. We used Markov Chain Monte Carlo sampling with a Metropolis-Hastings kernel, 10000 samples, and 1000 burn-in samples (i.e., additional sampling iterations before collecting samples), conditioning on the observations. Note that these only apply to Experiment 1 where agents have uncertainty about their own strength and their teammate's strength. In Experiments 2 and 3, subjects observe strength information and therefore those variables do not need to be sampled.

[5]For completeness, we show the predictions of all models with the Gini coefficient in the supplemental material (Figures S6-S10).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Our joint effort model includes two scaling parameters in the cost function: α, which denotes an agent's effort cost (laziness), and β, which determines the degree of penalization for unequal effort allocations. These were the only two free parameters in our model. In Experiment 1, both α and β are free parameters; we tuned them by hand and optimized them to maximize the correlation between model-predicted and empirically measured effort judgments pooled across participants. In Experiment 2 and Experiment 3, the α values were constrained by our experimental design (they were made explicit to the participants via the task description) and were not free parameters. To make model implementation consistent across experiments, we reused the β value from Experiment 1 for Experiment 2 and Experiment 3. See Table 1 for a complete list of α and β values used in the joint effort model. The three alternative models use the same α values as the joint effort model, but do not include the β parameter.

To assess how well the models match behavioral data, we calculated the Pearson correlation coefficient between model predictions and participants' judgments and plotted them against each other. Even though theoretically we could do likelihood-based model fitting, we are not making strong claims about the parametric details of the models. Rather, our claims concern the qualitative patterns of the model predictions.

3 Experiment 1

Our first experiment aims to test a few basic judgments about collaboration, including how people reason about competence and effort and how they predict how likely a collaborative event is to succeed. To be directly compatible with the model setup, we designed the experiment to be a game show where contestants attempt to lift a box by themselves or together with another contestant. We manipulated the individual lift outcome of different contestants shown to the participants and asked them to report judgments regarding what they predict would happen in a group lifting. We deliberately chose to not involve participants as active agents, but rather have them observe agents' behavior. This is because our work focuses on people's theory of mind, which is how people think *others* behave: Participants observe two agents' behavior and make judgments regarding how they think the agents collaborate.

Across different scenarios, we held the weight of the box constant during the group lift events, and we manipulated the reward for a successful lift and outcomes of agents' lifts in prior events to affect observers' beliefs about each agent's competence. Participants observed that some contestants refused to lift a box for a low reward, but readily lifted it for a higher reward. Beyond simply tracking agents' competence, these covariation data also provide information about how different agents respond to incentives. The different models make the following predictions:

- Strength: We predict that people are able to estimate contestants' strength based on observations of individual lift events. If a contestant lifts a box successfully, then they should be stronger than another contestant who fails to lift the same box. Furthermore, if they could lift the box with lower reward, then they should be stronger than contestants who require a higher reward, because accepting a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

lower reward indicates that less effort is needed to do the lifting. The joint effort model and the alternative models make similar predictions.

- Effort: When both contestants are strong enough to lift the box by themselves, we predict that people should expect them to exert less effort when they attempt to lift the box together. We also expect the models to make qualitatively different predictions. In particular, the joint effort model predicts this decrease in effort. The solitary effort model predicts no change in effort, since it assumes no difference between lifting the box themselves and lifting with another contestant. The compensatory effort model assumes that every contestant should expect the other one, who is strong enough to lift the box themselves, to be solely responsible for the lifting. As a result, it predicts that neither contestant will exert any effort. Finally, the maximum effort model predicts that the effort does not change from individual lifting to group lifting, since it assumes that contestants are always exerting all their effort.
- Lift probability: We predict that people should judge a group lifting to be more likely to succeed when the contestants involved are stronger. And when at least one contestant is strong enough to lift the box themselves, the group lift probability should be high. This pattern is predicted by the joint effort model, solitary effort model, and maximum effort model. In contrast, the compensatory effort model predicts that group lifting is impossible when at least one contestant is strong enough to lift the box by themselves. This is because the other contestant would expect the contestant who is strong enough to do the lifting and conversely, that contestant would expect the other one to exert their full effort, so they would reduce their effort accordingly, resulting in failure.

  We also predict that even if neither of the contestants is strong enough to lift the box themselves, it would still be possible for them to lift the box together. This prediction is consistent with the joint effort model and maximum effort model, which argue that “two are better than one.” The compensatory effort model predicts something similar in this case, because it is possible that the sum of the contestants’ strength equal the box weight, and if each contestant considers the remaining part after subtracting the other contestant’s strength requiring their full effort, that would mean both contestants exerting all their effort and possibly lifting the box successfully. The solitary effort model in this case predicts that group lifting is impossible, because neither contestant can lift the box themselves.

3.1 Materials and Methods

3.1.1 Participants—We recruited 50 participants from Amazon Mechanical Turk. Participants’ demographic information was not collected. Participants completed a comprehension check before they moved on to the experiment. They were not allowed to proceed until they answered all the comprehension check questions correctly. Participants received a base pay of $2 and a potential bonus payment up to $1. The amount of bonus they received was equal to the probability they put on the realized lift outcome on a randomly

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

picked round. We chose this bonus scheme because it is incentive compatible: Expected bonus is maximized by reporting their best estimate of the lift probability. To ensure data quality, we included two attention check questions in the experiment. Participants who failed one attention check were warned immediately to pay closer attention. Participants who failed both attention checks were asked to leave the experiment and they were not counted towards the 50 participants we recruited. A total of 10 participants failed one attention check, and we did not exclude their data in our analysis. The Harvard Institutional Review Board approved the experimental procedures and participants provided informed consent prior to the experiment.

3.1.2 Procedure—Participants observed six contests between different pairs of contestants (see Table 2 for a description of the contests; the order was randomized). In each contest, the contestants were given three attempts to lift a box, corresponding to three rounds. In the first two rounds, the contestants tried lifting the box themselves. The reward for lifting the box was $10 in Round 1 and $20 in Round 2. In the third round of each contest, the two contestants tried to lift the box together for a reward of $20 for each. Participants first saw the lift outcome of Round 1 and made strength judgments (1-10; 1 means extremely weak and 10 means extremely strong) and effort judgments (0-100%) for each contestant. For Round 2 and Round 3, they predicted the probability of the contestants lifting the box (0-100%) before seeing the outcome, then observed the actual outcome and made strength and effort judgments. Note that participants made effort judgments only when the outcome was Lift. Participants were informed that the weight of the box was always the same and equivalent to a strength of 5 (i.e., an average contestant with strength 5 exerting all of their effort would be able to lift the box). Participants also saw a table of all the previous outcomes when making their guesses. Figure 2 shows an illustration of the task.

3.1.3 Transparency and openness—Data and code for this and subsequent experiments are available at https://github.com/yyyxiang/competence_effort.

3.2 Results

As a preliminary check, we plotted participants' strength judgments on top of the model predictions (Figure 3A). The model predictions were similar to the behavioral data overall, except that the compensatory effort model and maximum effort model could not generate predictions in a few scenarios and rounds. The compensatory effort model could not predict contestants' strength in Round 3 of 'F,L;F,L', 'F,F;L,L', 'F,L;L,L', and 'L,L;L,L', since it predicts that the Round 3 outcome of these scenarios should always be Fail, but the observation is Lift. The maximum effort model could not make predictions for 'F,F;F,L', 'F,F;L,L', and 'F,L;L,L', simply because if contestants always exert 100% of their effort, then they would not fail in Round 1 and lift the box in Round 2. Note that the solitary effort model's predictions in all the scenarios except 'F,F;F,F' are disguised by that of joint effort model. Both the joint effort model and solitary effort model fail to make predictions for strength in Round 3 of 'F,F;F,L', as they both predict that the outcome should be Lift, conflicting the observation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

One major hypothesis concerns how effort judgments change from Round 2 to Round 3. If participants believed that contestants put in less effort when they worked together to lift the box in Round 3, compared to trying to lift the box by themselves in Round 2, then we should see a decrease in effort from Round 2 to Round 3. We selected trials from scenarios 'F,F;L,L', 'F,L;L,L', and 'L,L;L,L' where participants reported Round 2 effort judgments for both contestants, excluded Round 1 trials, and ran a linear mixed-effects regression regressing Effort on Round and Agent, with random effects for the intercept, Round, and Agent grouped by participants. Indeed, the regression results revealed that effort decreased from Round 2 to Round 3 [$t(49.0) = -5.160$, $p < .0001$]. Figure 3B visualizes this effect and shows that only the joint effort model makes this prediction. The solitary effort model predicts that effort should not change from Round 2 to Round 3, when contestants switched from lifting the box themselves to lifting together. The compensatory effort model could not predict contestants' effort in Round 3 and the maximum effort model could not make predictions for 'F,F;L,L', and 'F,L;L,L', as explained above.

Another hypothesis concerns the lift probability in Round 3. The lift probability should increase as contestants get stronger (moving from left to right along the x-axis). The lift probability should also be pretty high when at least one contestant had a successful lift, that is, all the scenarios except for 'F,F;F,F'. However, in 'F,F;F,F' when both contestants failed in both rounds, the lift probability should be non-zero, given that the two contestants were attempting to lift the box together. One-sample t-test showed that participants believed that the lift probability for 'F,F;F,F' was different from zero [$t(49) = 10.42$, $p < .0001$]. Figure 3C confirms these hypotheses and shows that the joint effort model is the only model that exhibits these patterns. The solitary effort model predicts that the lift probability is zero in 'F,F;F,F'. Again, the maximum effort model could not make predictions for 'F,F;F,L', 'F,F;L,L', and 'F,L;L,L'.

Figure 4 compares model predictions to participants' judgments regarding the lift probability, effort, and strength. Overall, the joint effort model provides the best fit to the behavioral data. Aside from missing predictions in certain scenarios and rounds (the compensatory effort model could not predict effort and strength in Round 3 for any scenarios except 'F,F;F,F', and the maximum effort model could not make predictions for scenarios 'F,F;F,L', 'F,F;L,L', and 'F,L;L,L'), the compensatory effort model failed to predict the lift probability. The solitary effort model and the maximum effort model failed for effort judgments.

3.3 Discussion

In Experiment 1, we studied participants' judgments of strength, effort, and lift probability in both individual lifting and collaborative lifting, based on observations of previous lift events. The joint effort model provides predictions that are qualitatively similar to the behavioral data, including the decrease in effort from Round 2 to Round 3, the non-zero lift probability in Round 3 of scenario 'F,F;F,F', and the high lift probability in Round 3 of all the other scenarios. The joint effort model also provides the overall best fit, compared to the alternative models. We conclude that people employ an intuitive theory of beliefs, desires, and competences, to reason recursively about joint effort.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4 Experiment 2

To test our theory's flexibility, in Experiment 2, we ask how people assign incentives to teams, based on knowledge about their strength. To differentiate between the models, we designed the experiment such that across multiple rounds in each contest, the strongest contestant is always involved in the group lifting. We also made sure that all the contestants in this experiment are strong enough to lift the box by themselves.

We expect qualitatively different predictions from the models. The joint effort model predicts a decrease in incentive when the contestants are stronger. The solitary effort model predicts no change in incentive across different rounds, because it assumes that the stronger contestant is effectively the one who determines the lower bound of the incentive required, and since that contestant is not changing across rounds in a contest, the incentive should not change either. The compensatory effort model and maximum effort model both predict zero incentive, though for different reasons. The compensatory effort model predicts that since every contestant is strong enough to lift the box by themselves, two contestants would never lift the box together successfully, thus no incentive should be wasted on this impossible mission. The maximum effort model, in contrast, predicts that the contestants would lift the box regardless of the incentive, assuming that everyone would always be exerting 100% of their effort. Therefore, $0 incentive would be the best choice.

4.1 Materials and Methods

4.1.1 Participants—We recruited 98 participants from Amazon Mechanical Turk. Participants' demographic information was not collected. Same as in Experiment 1, participants completed a comprehension check before starting the main experiment. Participants received a base pay of $0.5 and a potential bonus payment up to $6. The bonus payment depended on the incentive participants provided to the contestants and whether the lifting turned out successful given the incentive. Specifically, for every successful lift, participants received $0.4; for every dollar of incentive they gave out, $0.002 was deducted from their bonus payment, regardless of the lift outcome. [6] The Harvard Institutional Review Board approved the experimental procedures and participants provided informed consent prior to the experiment.

4.1.2 Procedure—Participants observed five different contests between different teams of contestants. In each contest, participants saw four different contestants and the minimum incentive each would accept to lift the box alone as a reference point. To give participants a better sense of what the minimum incentives meant, we converted the minimum incentive each contestant required to an estimate of their strength (see Figure 5 for an illustration of the task). In each round of the contest, the strongest contestant was always among the pair of contestants attempting to lift the box, while the other contestant was one of the remaining contestants. We intentionally included a contestant (Contestant D) who is as

[6]Note that we did not use attention checks in Experiments 2 and 3. Our results from Experiment 1 suggest that participants' overall performance on the attention checks was high (3 participants failed the first attention check, 7 participants failed the second attention check), and that the interpretation of the results does not change regardless of whether we exclude participants who failed one attention check.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

strong as the strongest contestant (Contestant A). This is when the models (specifically, the joint effort model and the solitary effort model) make the most distinct incentive predictions quantitatively: The solitary effort model's predictions do not change as long as Contestant D is not stronger than Contestant A, whereas the joint effort model's predictions decrease when the teammate gets stronger. This difference in prediction is the largest when the teammate is as strong as the strongest contestant. Participants decided how much incentive to provide each pair of contestants to lift the box together. They were told that the same incentive would go to both contestants, i.e., if participants provided $50 to a pair of contestants, then both contestants would get $50 if they lifted the box together successfully. The box weight was fixed at 5, as in Experiment 1. Participants also completed three training trials where they tested how a random contestant would respond to different incentives.

4.2 Results

To test the hypothesis that the incentive will decrease when one teammate becomes stronger, we constructed a linear mixed-effects model, in which we regressed Incentive on the strength of the two contestants involved in the lifting, with participant-level random effects for all the regressors. We found that controlling for the strength of the stronger contestant who remained the same throughout each contest, the strength of the weaker contestant had a statistically significant negative effect on the incentive provided by the participants [$t(1167.0) = -12.595$, $p < .0001$]. This result is visualized in Figure 6A.

Both the compensatory effort and maximum effort models predict that the optimal incentive should be $0 across all rounds and contests. The solitary effort model predicts that the incentive provided to the contestants should remain the same when the stronger contestant was unchanged, as shown in Figure 6A. Only the joint effort model showed the trend of incentive decreasing within each contest. This is also validated by the high Pearson correlation coefficient ($r = 0.95$; Figure 6B).

4.3 Discussion

Experiment 2 is the application of our theory to the problem of incentive selection. With a task that controlled for the stronger contestant's strength, the joint effort model showed qualitatively similar predictions to the behavioral data, while all the other models made distinct predictions regarding how the incentive should change as a function of teammate strength.

5 Experiment 3

In Experiment 3, we further extend the model to the problem of team selection. Besides manipulating contestants' strength, we manipulated their effort cost (the α parameter in our model). We also designed different box weights such that weaker but more hard-working contestants and stronger but lazier contestants are favored differently: Weaker but more hard-working contestants should be preferred when the box is light, due to their lower effort cost. In contrast, stronger but lazier contestants should be preferred when the box is heavy because weaker contestant would not be able to lift a box that is too heavy for them no matter how hard they try. In addition, if we allow contestant-specific rewards (i.e., different

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

contestants on the team can receive different rewards) then we should see that the incentive allocated to each contestant correlates with how often they are selected.

The models generate distinct predictions regarding who to select. The joint effort model follows a similar logic as above. The solitary effort model will always pick contestants who are able to lift the box themselves, and if forced to include another contestant on the team, it will not expect the other contestant to exert any effort. The compensatory effort model only considers a group lifting possible when the sum of the contestants' strength equal the box weight (i.e., when both contestants exert all their effort). Otherwise, it assumes that the group lifting is impossible anyway and therefore values every contestant equally. The maximum effort model makes similar predictions as the joint effort model in terms of who to select, but when more than one group of contestants are able to succeed in lifting, it values them all equally, since it doesn't take into account their effort costs; all it cares is each contestant's strength. These differences should also be reflected in the incentives assigned to each contestant.

5.1 Materials and Methods

5.1.1 Participants—We recruited 101 participants from Amazon Mechanical Turk. Participants' demographic information was not collected. Same as in Experiment 1 and Experiment 2, participants completed a comprehension check before starting the main experiment. Participants received a base pay of $0.5 and a potential bonus payment up to $3. The bonus payment depended on the incentive participants provided to the contestants and whether the lifting was successful given the incentive: For every successful lift, participants received $1; for every dollar of incentive they gave out, $0.01 was deducted from their bonus payment, regardless of the lift outcome. The Harvard Institutional Review Board approved the experimental procedures and participants provided informed consent prior to the experiment.

5.1.2 Procedure—Participants observed three different contests between three contestants who had different strength and laziness. Participants knew each contestant's strength and the minimum incentive each would accept to lift the heaviest box they could (box weight equivalent to their strength), as shown in Figure 7A. This provides participants information about each agent's effort cost, which equals the minimum incentive they would accept to lift the heaviest box they could (i.e., minimum incentive is equivalent to $a \cdot 100\%$). The box weight varied from contest to contest, ranging from 10 to 7 to 6. In each contest, participants first selected two contestants to do the lifting together, then decided the incentive they would provide to each contestant. Note that in this experiment, each contestant's incentive was contestant-specific. In other words, instead of both contestants receiving the same incentive, participants chose a separate incentive for each contestants. We constrained participants' budget in each contest to $50 (i.e., the total incentive they gave out in each contest could not exceed $50). Participants also completed six training trials where they tested how a random contestant would respond to different incentives given two different box weights.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5.2 Results

When the box weight was 10, the joint effort model predicts that Contestant A will always be selected, because without Contestant A, Contestant B and Contestant C's total strength was smaller than the box weight. As for the other teammate, the joint effort model predicts that Contestant C will be preferred over Contestant B, because selecting Contestant B would entail a higher effort cost. When the box weight was 6 or 7, however, the joint effort model predicts that Contestant B and Contestant C will be selected because their total strength was equal to or exceeded the box weight, and their effort would cost less than Contestant A. This was indeed what we saw from the behavioral data (Figure 7B): When the box weight was 10, 92.08% of the participants ($n = 93$) selected Contestant A. More participants selected Contestant C (64.36% of the participants, $n = 65$) over Contestant B (43.56% of the participants, $n = 44$). When the box weight was 7, only 35.64% of the participants ($n = 36$) selected Contestant A. 89.11% of the participants ($n = 90$) selected Contestant B, and 75.25% of the participants ($n = 76$) selected Contestant C. When the box weight was 6, only 25.74% of the participants ($n = 26$) selected Contestant A. 93.07% of the participants ($n = 94$) selected Contestant B, and 81.19% of the participants ($n = 82$) selected Contestant C. As shown in Figure 7B, the joint effort model is the only model that shows a pattern qualitatively similar to the behavioral data. Figure 8 validates this conclusion with a Pearson correlation coefficient of 0.91 between the behavioral data and the joint effort model's predictions.

We also looked at how the total incentive and individual incentive changed with box weight. If the participants believed that the box was liftable in all three contests, then the total incentive should increase with box weight. As expected, we saw a monotonic increase in the total incentive allocated to contestants as the box weight increased (Figure 7C), confirmed by a linear mixed-effects regression which regressed the total incentive on the box weight, with random effects of the intercept and the box weight grouped by participants [$t(201.0) = 12.232$, $p < .0001$].

The joint effort model predicts that the incentive allocated to each contestant should change according to how likely they are to be selected. Across the three contests, we should see an overall trend of Contestant A's incentive increasing with box weight, and Contestant B and Contestant C's incentive decreasing with box weight. There should not be a big difference between box weight of 6 and box weight of 7, so the transition from box weight of 7 to box weight of 10 should primarily determine the trend. From three separate linear mixed-effects regressions constructed for the three contestants, each of which regressed the incentive allocated to the contestant on the box weight, with random effects of the intercept and box weight grouped by participants, we found that the incentive for Contestant A increased with box weight [$t(201.0) = 14.801$, $p < .0001$], incentive for Contestant B decreased with box weight [$t(229.2) = -10.56$, $p < .0001$], and incentive for Contestant C decreased with box weight [$t(201.0) = -1.910$, $p = .058$]. The incentive for Contestant C showed a non-significant trend towards decreasing with box weight, likely because Contestant C's minimum incentive is too low to allow for sufficient variation. These effects are visualized in Figure 7D.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

To confirm these results, we directly tested the correlation between the incentive allocated to each contestant and whether they were selected. From three separate linear mixed-effects regressions (one for each contestant) with incentive as the dependent variable and a categorical variable that indicates whether the contestant was selected being the predictor variable, along with random effects of the intercept and the categorical variable grouped by participants, we found that incentive increased when the contestant was selected. This effect was significant for all three contestants: Contestant A [$t(105.8) = 35.15$, $p < .0001$], Contestant B [$t(176.9) = 31.82$, $p < .0001$], and Contestant C [$t(102.1) = 9.544$, $p < .0001$].

5.3 Discussion

Experiment 3 asked participants to make judgments regarding team selection: who to select and how much incentive for each teammate. As predicted by the joint effort model, participants cared about the total strength of the contestants when deciding whether they would be able to lift the box together, therefore favored the strongest contestant (Contestant A) when the box was very heavy (box weight = 10). When the box was lighter (box weight = 6 or 7), however, participants favored contestants who were more hard-working, which entailed lower effort cost (Contestants B and C). The total incentive increased with box weight and the incentive participants provided to each contestant corresponded to their selection of teammates. For all three judgments, the joint effort model's predictions were qualitatively similar to the behavioral data, while the other three models all made predictions that deviated systematically from the data.

From Figure 7B, we see that humans exhibited a preference for certain contestants over others; for example, when the box weight was 6 and 7, some participants selected Contestant A over Contestant C, thereby showing a clear preference for Contestant B over Contestant C. This pattern was not observed in the joint effort model's predictions. It is worth noting, however, that there is an important methodological difference between Experiment 3 and Experiment 1. In Experiment 3, each contestant's strength is determined by the experimental setup; therefore the model does not infer contestants' strength, which makes it impossible for the joint effort model to choose probabilistically. In other words, the joint effort model cannot choose another partner because of the way that the decision rule is set up. It is possible that, with a different decision rule, we could bring the quantitative pattern of the data arbitrarily close to human judgments, but that would not change the substance of the model. The important point is that the joint effort model captures human intuitions about which collaborative partner would be preferred.

6 General Discussion

We presented a belief-desire-competence theoretical framework which extends theory of mind reasoning to include an agent-specific representation of competence and effort. We applied this framework to collaborative cognition, where multiple agents work together in order to achieve goals that may be impossible for any individual. To succeed in these collaborative tasks, we argued that agents must know what other agents are capable of doing and how much effort they are willing to exert.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Through three studies, we demonstrated that people make judgments about competence, effort, incentives, and team structure that are in broad agreement with our framework. In Experiment 1, we found that people can infer others' competence through observations of behavior. More importantly, people can generalize individual agents' behavior to multi-agent contexts and make inferences about group behavior, including predicting whether a collaboration will be successful, and inferring the amount of effort collaborators exert, without any direct observation. In Experiment 2, we found that people assigned incentives to teams based on considerations of their strength, and in Experiment 3, we found that people allocated individuals to teams and selected incentives for them according to agent-specific attributes and agent-general tasks. Taken together, we showed that people are capable of making inferences regarding collaborative behavior and applying these inferences to tackle novel problems of incentive selection and team selection.

We showed that human judgments of collaborative behavior were well-predicted by a multi-agent joint effort model grounded in recursive reasoning of effort between agents. Alternative models that did not invoke recursive reasoning failed to capture the data, suggesting that recursive theory of mind plays a central role in human collaboration.

While the joint effort model qualitatively resembled the data, people reported lower probabilities, greater effort, and higher incentives compared to the joint effort model. This raises the possibility that people are risk-averse, erring on the side of safety by preferring a slightly lower bonus over no bonus. We explored this possibility by creating a *safe joint effort model,* in which a hindrance factor is added to hedge the model's predictions, as we speculate that participants are doing. Instead of satisfying Eq. 4, the safe joint effort model requires that the sum of agents' effort times strength has to be greater than or equal to the box weight plus this additional hindrance factor. This model yields predictions that are quantitatively closer to human judgments. Since the predictions of this safe joint effort model are qualitatively similar to the original joint effort model, we decided for exposition purposes to keep the simpler model in the main text but include the elaborated model in the supplemental material (Figures S1-S5). The safe joint effort model is also able to offer a plausible explanation regarding why participants in Experiment 1 were able to make strength judgments in Round 3 of scenario 'F,F;F,L' while the joint effort model could not: When the hindrance factor is large (greater than 1.5), the safe joint effort model is able to make predictions for this scenario.

Our work complements past research on collaboration. Evolutionary theories of collaboration address how collaboration could have emerged as evolutionarily stable strategies at the population level (Henrich and Muthukrishna, 2021). Economic theories of collaboration are concerned with the structure of the task: under what combinations of incentives agents could be expected to cooperate (Lopes, 1994). The machine learning literature on collaboration focuses on how agents carry out sub-tasks in cooperative multi-agents systems (e.g., Oliehoek and Amato, 2016). Our work provides a different perspective on collaboration, grounded in a formal description of the cognitive capacities that enable individuals to coordinate collaborative actions, and validated through tasks with rich psychological structure.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

One limitation of our work is that we constrained the experimental setup to teams of two agents; in principle, however, our framework can generate predictions for larger teams. Moving forward, our framework may shed light not only on how humans apply belief-desire-competence reasoning to larger collaborations, but also what the boundaries of these reasoning abilities may be. From a computational perspective, we might expect agents to replace optimal recursive reasoning (which is intractable for large groups) with simpler heuristics (e.g. Golman et al., 2020). When viewed through a resource-rational lens (Gershman et al., 2015; Lieder and Griffiths, 2020), the heuristics used by the alternative models are plausible "shortcuts" to joint inference, and these models may be useful in diagnosing why collaborations sometimes fail. An open question for future work is how cognitively bounded groups of agents can realize resource-rational collaboration by optimally allocating their cognitive resources.

Understanding how well these inferences scale to larger teams—and when they begin to break down—may also provide a cognitive perspective on the design of organizations: Organizations make social reasoning easier by breaking down a large, sprawling structure into smaller work teams or chains of command, thus limiting the number of individuals with whom any one worker has to coordinate. Understanding the limitations of human social reasoning may inform the design of organizations and administrative structures that are adapted to those limitations (Kozlowski and Ilgen, 2006).

Another limitation of our work is that we operationalized competence as physical strength. Moreover, we simplified our analysis by treating strength as a static, scalar quantity. In real-world problems, competence is instead multidimensional and dynamic. Even within the domain of physical strength, one might be able to lift more weight during deadlifts than during chest presses, and the heaviest weight that one can lift may increase with consistent practice. Moving beyond the physical domain, real-world collaborations unite the efforts of people with varying cognitive skills and expertise, and the most relevant expertise that one has to offer may depend on the composition of the rest of the group—for example, an individual researcher with expertise in both computational modeling and experimental design may tackle the computational modeling in a team composed mostly of experimentalists, and implement experiments in a team of mostly theorists. Thus, one intriguing extension of our framework would be to allow agents to have varying degrees of competence in varying domains, effectively operationalizing competence as a vector rather than as a scalar quantity. In doing so, our framework may then be extended to predict how teams divide labor based on the expertise of agents in different tasks.

Our framework can be extended to treat competence as a dynamic, rather than a static, property. Competence can both increase and decrease over long periods of work. On one hand, after performing the same task many times—such as lifting a heavy weight, or manufacturing pins, or building computational models—individuals may become more efficient, able to carry out the same work for less effort. Indeed, foundational economic accounts suggest that this increase in competence is one of the chief benefits of division of labor (Smith, 2010): by specializing narrowly, individual workers can become more efficient at their particular task, thus outperforming generalists. On the other hand, there is also a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

downside to this repetition: Over time, an individual may grow bored with doing the same task, or carry out the same task less efficiently due to fatigue.

One particularly important problem for future work is understanding how skills are taught and transferred over long-term collaborations, such as in apprenticeships between craftspeople or in the mentoring relationship between an advisor and their student. Indeed, evolutionary accounts of teaching propose that teaching co-evolved with the use of complex tools, where more complex technical skills require more sophisticated, efficient means of transferring those skills between individuals (Lucas et al., 2020; Caldwell et al., 2018). However, little is known about how teachers select what skills to impart to their students, particularly in contexts where students can then hone these skills on their own through practice. Most laboratory experiments of teaching instead focus on how teachers select information that will change the mental states of a learner—such as what they believe (Shafto et al., 2014) or value (Ho et al., 2021)—rather than their skills (but see Kleiman-Weiner et al. 2020).

7 Conclusion

In sum, we have provided evidence that collaborative decisions are scaffolded by recursive inferences about others' effort and competence. We presented a belief-desire-competence framework that captures qualitative patterns in human judgments across a variety of decisions that are critical for collaboration, such as predicting whether a joint task will succeed, deciding what incentives to provide to collaborators, and choosing whom to recruit for a collaborative task. Moving forward, important tasks for future work include understanding the limits of these inferences—such as how well they scale to larger groups—and building dynamic, multidimensional representations of competence into the model. Ultimately, our goal is to more completely formalize the rich cognitive capacities that support human collaboration.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the Center for Brains, Minds and Machines (CBMM) and funded by an NSF STC award (award number CCF-1231216) to S.J.G. and an NIMH K00 award (award K00MH125856) to N.V. We thank Weiyao Dong for designing the task aesthetics.

References

Atkinson AB et al. (1970). On the measurement of inequality. Journal of economic theory, 2(3):244–263.

Baer C and Odic D (2022). Mini managers: Children strategically divide cognitive labor among collaborators, but with a self-serving bias. Child Development, 93(2):437–450. [PubMed: 34664258]

Baker CL, Jara-Ettinger J, Saxe R, and Tenenbaum JB (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behaviour, 1(4):1–10.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Baker CL, Saxe R, and Tenenbaum JB (2009). Action understanding as inverse planning. Cognition, 113(3):329–349. [PubMed: 19729154]

Blake P, McAuliffe K, Corbit J, Callaghan T, Barry O, Bowie A, Kleutsch L, Kramer K, Ross E, Vongsachang H, et al. (2015). The ontogeny of fairness in seven societies. Nature, 528(7581):258–261. [PubMed: 26580018]

Bratman ME (1992). Shared cooperative activity. The philosophical review, 101(2):327–341.

Caldwell CA, Renner E, and Atkinson M (2018). Human teaching and cumulative cultural evolution. Review of Philosophy and Psychology, 9(4):751–770. [PubMed: 30595765]

Campano F and Salvatore D (2006). Income Distribution: Includes CD. Oxford University Press.

Fehr E and Schmidt KM (1999). A theory of fairness, competition, and cooperation. The quarterly journal of economics, 114(3):817–868.

Gershman SJ, Horvitz EJ, and Tenenbaum JB (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. Science, 349:273–278. [PubMed: 26185246]

Golman R, Bhatia S, and Kane PB (2020). The dual accumulator model of strategic deliberation and decision making. Psychological review, 127:477–504. [PubMed: 31868393]

Goodman ND and Stuhlmüller A (2014). The design and implementation of probabilistic programming languages.

Grosz BJ and Kraus S (1996). Collaborative plans for complex group action. Artificial Intelligence, 86(2):269–357.

Hamann K, Warneken F, Greenberg JR, and Tomasello M (2011). Collaboration encourages equal sharing in children but not in chimpanzees. Nature, 476(7360):328–331. [PubMed: 21775985]

Henrich J and Muthukrishna M (2021). The origins and psychology of human cooperation. Annual Review of Psychology, 72:207–240.

Ho MK, Cushman F, Littman ML, and Austerweil JL (2021). Communication in action: Planning and interpreting communicative demonstrations. Journal of Experimental Psychology: General.

Jara-Ettinger J and Gweon H (2017). Minimal covariation data support future one-shot inferences about unobservable properties of novel agents. In CogSci.

Jara-Ettinger J, Gweon H, Schulz LE, and Tenenbaum JB (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. Trends in Cognitive Sciences, 20:589–604. [PubMed: 27388875]

Jara-Ettinger J, Gweon H, Tenenbaum JB, and Schulz LE (2015). Children's understanding of the costs and rewards underlying rational action. Cognition, 140:14–23. [PubMed: 25867996]

Karau SJ and Williams KD (1993). Social loafing: A meta-analytic review and theoretical integration. Journal of personality and social psychology, 65(4):681.

Khalvati K, Park SA, Mirbagheri S, Philippe R, Sestito M, Dreher J-C, and Rao RP (2019). Modeling other minds: Bayesian inference explains human choices in group decisionmaking. Science advances, 5(11):eaax8783. [PubMed: 31807706]

Kleiman-Weiner M, Sosa F, Thompson B, van Opheusden S, Griffiths T, Gershman S, and Cushman F (2020). Downloading culture. zip: Social learning by program induction. In CogSci.

Kozlowski SW and Ilgen DR (2006). Enhancing the effectiveness of work groups and teams. Psychological science in the public interest, 7(3):77–124. [PubMed: 26158912]

Latané B, Williams K, and Harkins S (1979). Many hands make light the work: The causes and consequences of social loafing. Journal of personality and social psychology, 37(6):822.

Leonard JA, Garcia A, and Schulz LE (2020). How adults' actions, outcomes, and testimony affect preschoolers' persistence. Child development, 91(4):1254–1271. [PubMed: 31502258]

Leonard JA, Lee Y, and Schulz LE (2017). Infants make more attempts to achieve a goal when they see adults persist. Science, 357(6357):1290–1294. [PubMed: 28935806]

Lieder F and Griffiths TL (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. Behavioral and Brain Sciences, 43:e1.

Lopes LL (1994). Psychology and economics: Perspectives on risk, cooperation, and the marketplace. Annual Review of Psychology, 45(1):197–227.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Lucas AJ, Kings M, Whittle D, Davey E, Happé F, Caldwell CA, and Thornton A (2020). The value of teaching increases with tool complexity in cumulative cultural evolution. Proceedings of the Royal Society B, 287(1939):20201885. [PubMed: 33203332]

Lucca K, Horton R, and Sommerville JA (2020). Infants rationally decide when and how to deploy effort. Nature human behaviour, 4(4):372–379.

Magid RW, DePascale M, and Schulz LE (2018). Four-and 5-year-olds infer differences in relative ability and appropriately allocate roles to achieve cooperative, competitive, and prosocial goals. Open Mind, 2(2):72–85.

Oliehoek FA and Amato C (2016). A Concise Introduction to Decentralized POMDPs. Springer.

Pollack ME (1990). Plans as complex mental attitudes. Intentions in communication, 77(104):277–282.

Premack D and Woodruff G (1978). Does the chimpanzee have a theory of mind? Behavioral and brain sciences, 1(4):515–526.

Regidor E (2004). Measures of health inequalities: part 1. Journal of epidemiology and community health, 58(10):858. [PubMed: 15365113]

Sadras V and Bongiovanni R (2004). Use of lorenz curves and gini coefficients to assess yield inequality within paddocks. Field Crops Research, 90(2-3):303–310.

Searle JR (1990). Collective intentions and actions john r. searle. Intentions in communication, 401.

Sebanz N and Knoblich G (2021). Progress in joint-action research. Current Directions in Psychological Science, 30(2):138–143.

Sen A, Sen MA, Foster JE, Amartya S, Foster JE, et al. (1997). On economic inequality. Oxford university press.

Shafto P, Goodman ND, and Griffiths TL (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cognitive psychology, 71:55–89. [PubMed: 24607849]

Smith A (2010). The Wealth of Nations: An inquiry into the nature and causes of the Wealth of Nations. Harriman House Limited.

Tabibnia G, Satpute AB, and Lieberman MD (2008). The sunny side of fairness: preference for fairness activates reward circuitry (and disregarding unfairness activates self-control circuitry). Psychological science, 19(4):339–347. [PubMed: 18399886]

Thomas V, Wang Y, and Fan X (2001). Measuring education inequality: Gini coefficients of education, volume 2525. World Bank Publications.

Tomasello M, Carpenter M, Call J, Behne T, and Moll H (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and brain sciences, 28(5):675–691. [PubMed: 16262930]

Tomasello M and Hamann K (2012). The 37th sir frederick bartlett lecture: Collaboration in young children. Quarterly Journal of Experimental Psychology, 65(1):1–12.

Török G, Pomiechowska B, Csibra G, and Sebanz N (2019). Rationality in joint action: Maximizing coefficiency in coordination. Psychological science, 30(6):930–941. [PubMed: 31088200]

Török G, Stanciu O, Sebanz N, and Csibra G (2021). Computing joint action costs: co-actors minimize the aggregate individual costs in an action sequence. Open Mind, pages 1–13. [PubMed: 34485794]

Vesper C, Butterfill S, Knoblich G, and Sebanz N (2010). A minimal architecture for joint action. Neural Networks, 23(8-9):998–1003. [PubMed: 20598504]

Williams KD and Karau SJ (1991). Social loafing and social compensation: The effects of expectations of co-worker performance. Journal of personality and social psychology, 61(4):570. [PubMed: 1960649]

Wu SA, Wang RE, Evans JA, Tenenbaum JB, Parkes DC, and Kleiman-Weiner M (2021). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. Topics in Cognitive Science, 13(2):414–432. [PubMed: 33829670]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

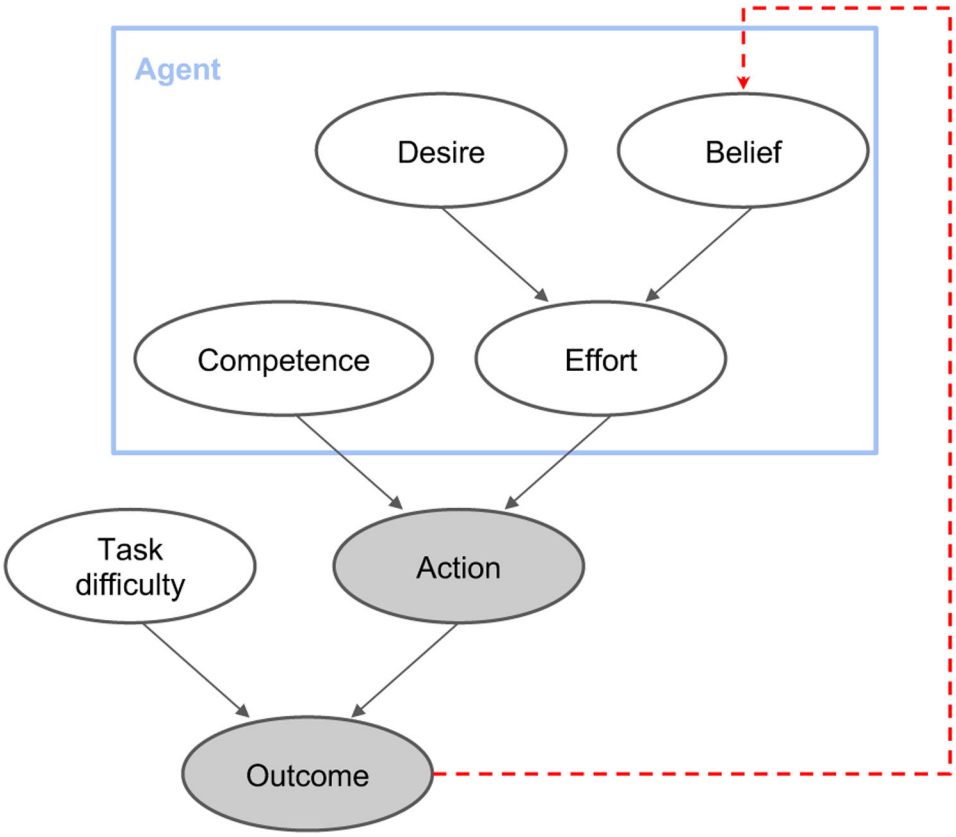

Figure 1:
An illustration of the belief-desire-competence framework. Shaded nodes represent observed variables; unshaded nodes represent latent variables to be inferred. The agent's desires and beliefs determine their effort. Their competence and effort together determine the action to take. The outcome is a function of their action and the task difficulty. Observing the outcome updates the agent's beliefs.

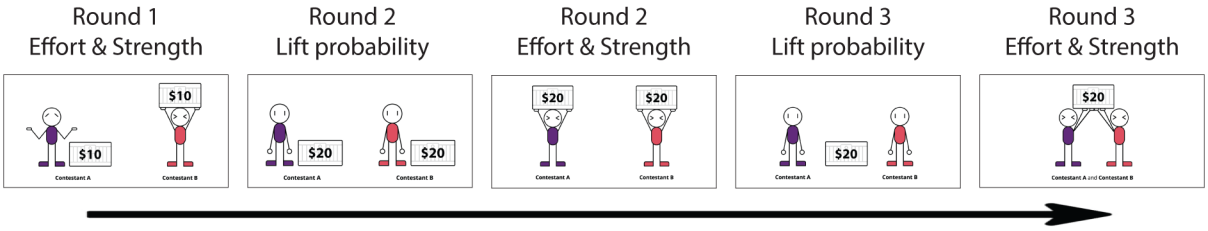

Figure 2:

Example contest in Experiment 1. Each contest consists of three rounds. In Round 1, participants observed individual lift outcomes and reported their judgments of contestants' effort and strength. In Round 2, participants first guessed the lift probability when reward is increased from $10 to $20, then reported effort and strength judgments after observing Round 2 outcomes. In Round 3, participants guessed the probability of the two contestants lifting the box together, and reported effort and strength judgments of each contestant after observing the Round 3 outcome. At all times, participants saw a table of previous lift outcomes in that contest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

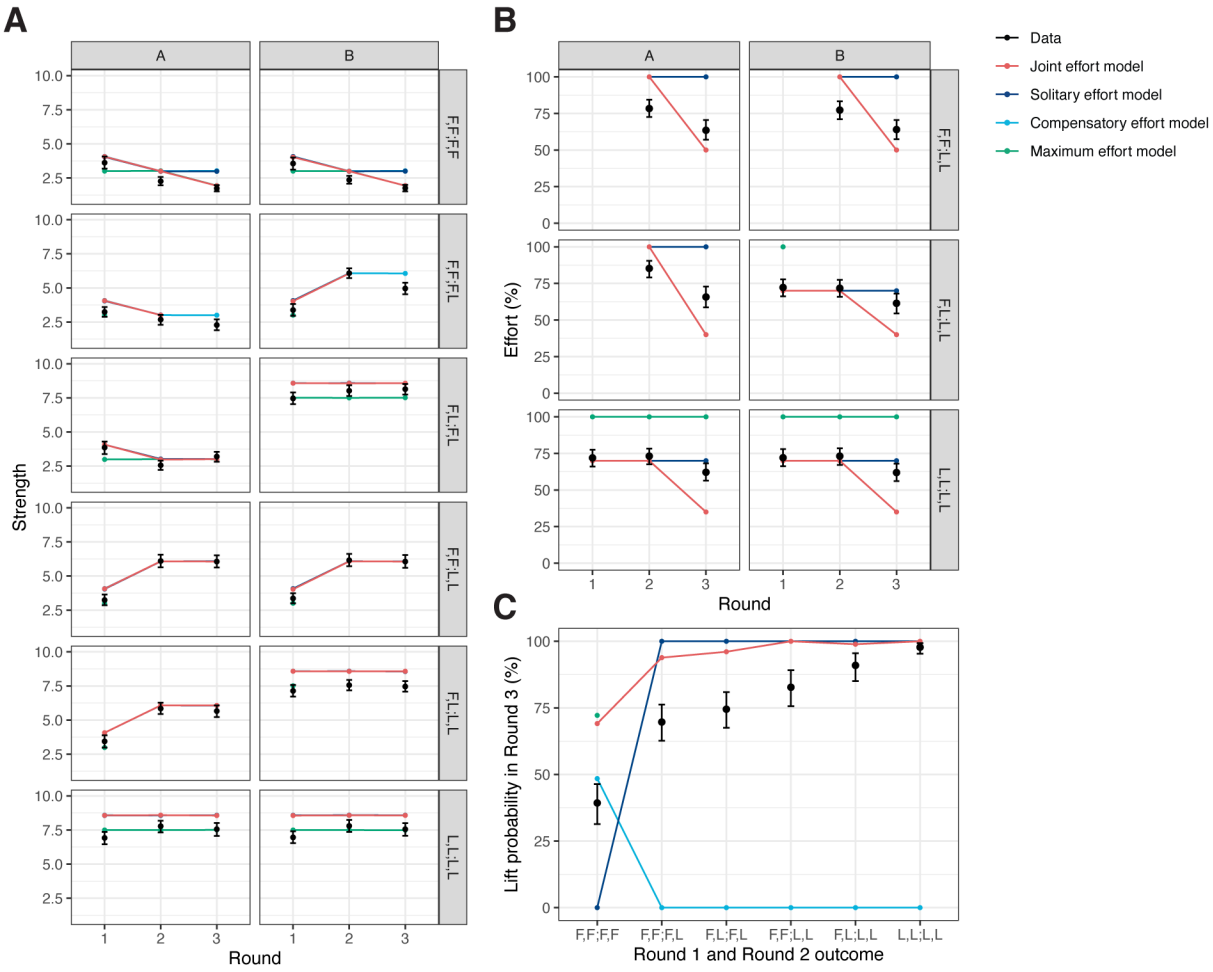

Figure 3:

(A) Predictions of contestants' strength in Experiment 1. (B) Predictions of contestants' effort in Experiment 1. Effort judgments were not elicited when the lift outcome was Fail. (C) Predictions of the lift probability in Round 3 of Experiment 1. Model simulations averaged over 10 runs. Error bars indicate bootstrapped 95% confidence intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

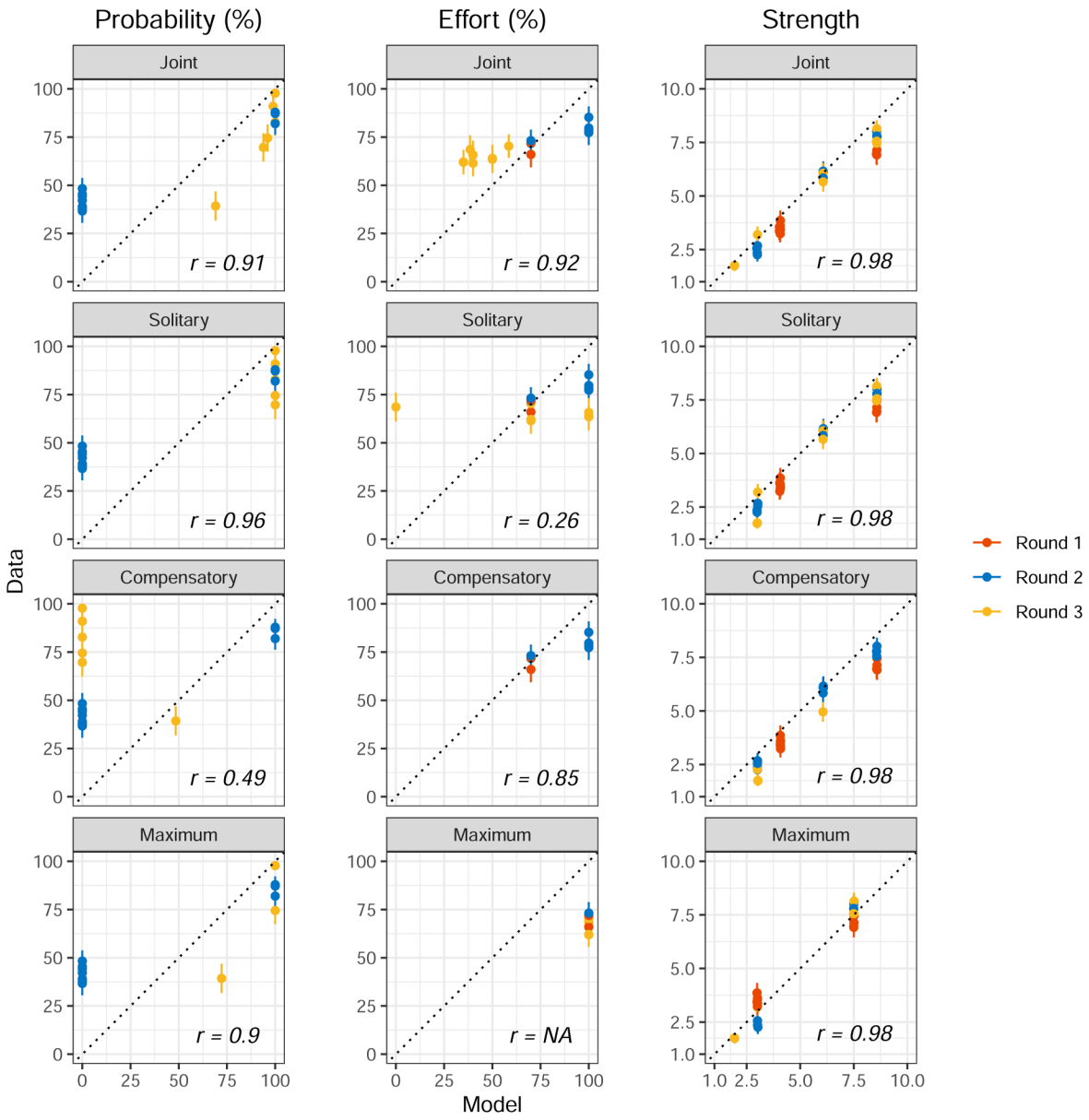

Figure 4:

Comparing model predictions to behavioral data across predictions of lift probability, effort, and strength in Experiment 1. Model simulations averaged over 10 runs. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars indicate 95% normal confidence intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Figure 5:

An example contest in Experiment 2. Each contest contains three rounds, and in each round, the strongest contestant (to the left) attempts to lift a box with one of the three contestants to the right. Participants reported how much incentive they were willing to provide to each pair of contestants.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

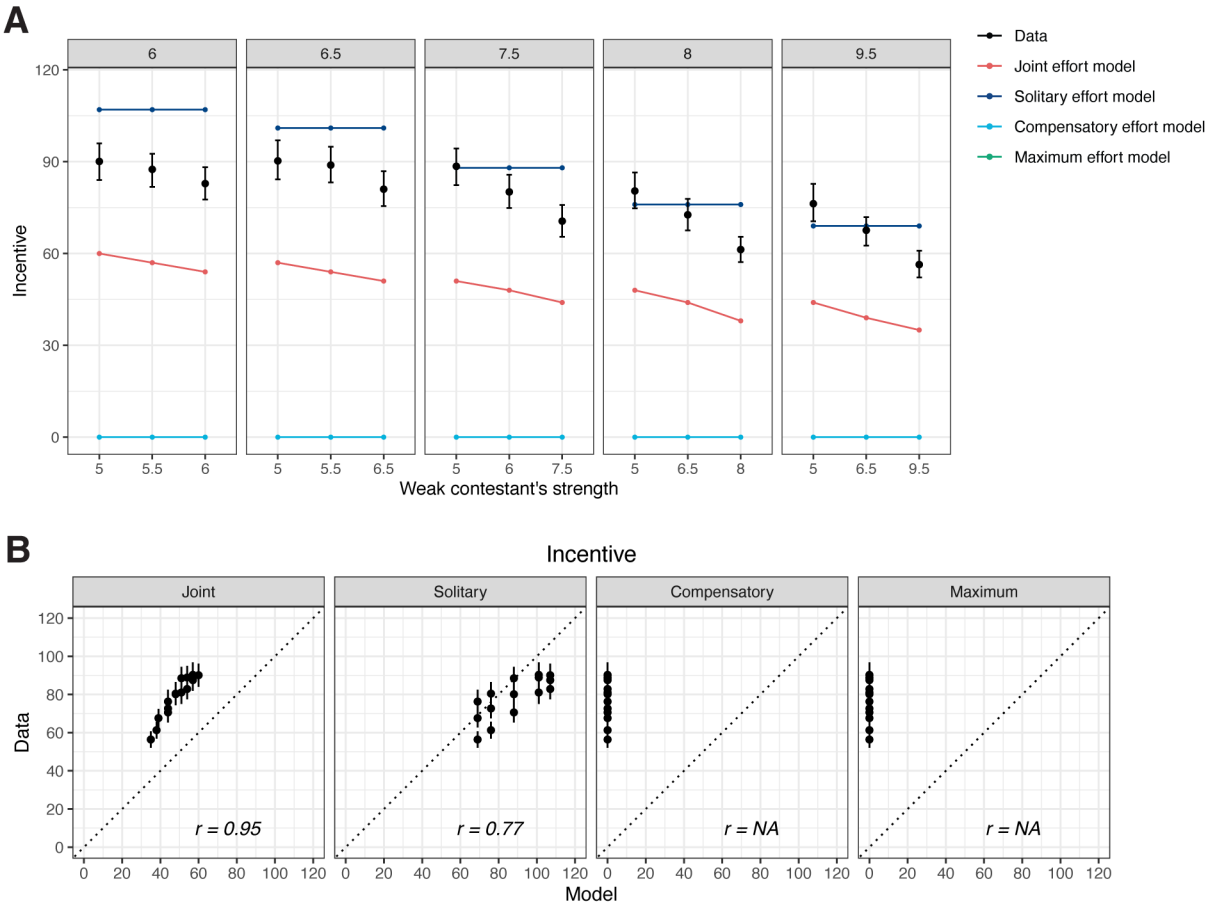

Figure 6:

(A) Incentive provided to pairs of contestants in Experiment 2. Each panel corresponds to the strongest contestant's strength in each contest. (B) Comparison between behavioral data and model predictions of incentive in Experiment 2. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars in (A) indicate bootstrapped 95% confidence intervals; error bars in (B) indicate 95% normal confidence intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

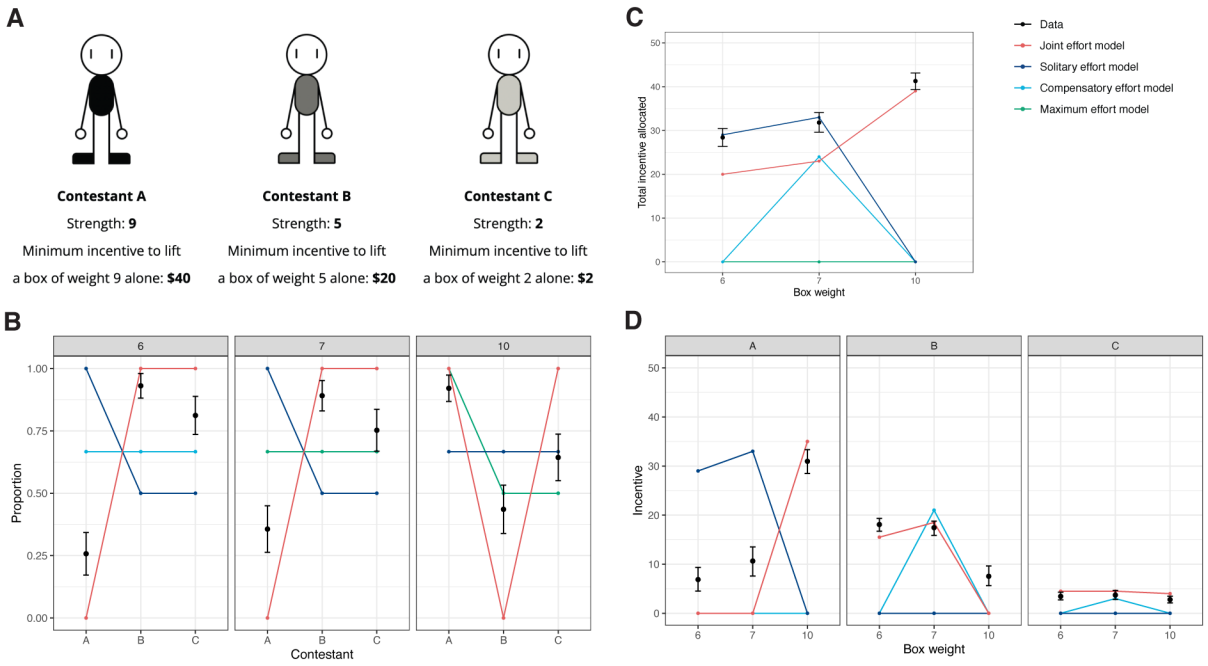

Figure 7:

(A) Contestants in Experiment 3. (B) Proportion of times each contestant was selected for each contest in Experiment 3. Contests are identified by their box weights, which appear above each plot. (C) Total incentive provided to contestants in Experiment 3. (D) Incentive provided to each contestant in Experiment 3. Each plot corresponds to an individual contestant. Error bars in (B) indicate 95% confidence intervals of proportions; error bars in (C) and (D) indicate bootstrapped 95% confidence intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

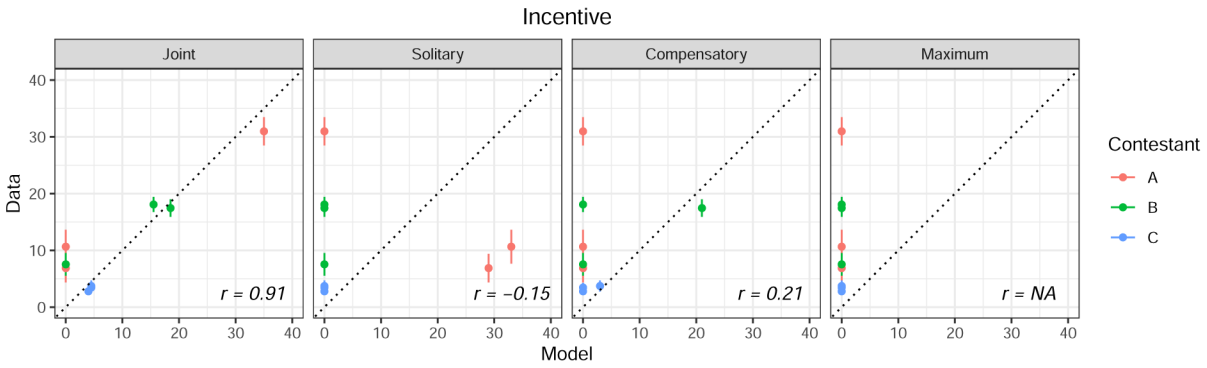

Figure 8:

Comparison between behavioral data and model predictions of incentive in Experiment 3. Pearson correlation coefficient shown at the bottom right of each subplot. Error bars indicate 95% normal confidence intervals.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1:

Scaling parameter values.

	α	β
Experiment 1	13.5	24.5
Experiment 2	125	24.5
Experiment 3	Agent-specific (40, 20, or 2)	24.5

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2:

Lift outcome Experiment 1 scenarios.

Scenario	Agent	Round 1	Round 2	Round 3
F,F;F,F	A	Fail	Fail	Fail
	B	Fail	Fail	
F,F;F,L	A	Fail	Fail	Fail
	B	Fail	Lift	
F,L;F,L	A	Fail	Fail	Lift
	B	Lift	Lift	
F,F;L,L	A	Fail	Lift	Lift
	B	Fail	Lift	
F,L;L,L	A	Fail	Lift	Lift
	B	Lift	Lift	
L,L;L,L	A	Lift	Lift	Lift
	B	Lift	Lift	

Note: We re-organized the data such that Agent A is the weaker contestant and Agent B is the stronger contestant in each contest when they had different outcomes. The side was randomized in the experiment.